2018 年度
早稲田大学大学院基幹理工学研究科情報理工・情報通信専攻
修士論文

言語情報を活用したシーングラフ生成

黒 澤　郁 音
(5117f027-1)

提出日：2019.02.01
主指導：林良彦教授
指導教員：小林哲則教授
研究指導名：知覚情報システム研究

目 次

表 目 次

図 目 次

概要

シーングラフ生成を高精度かつ効率的に行うための手法の提案を行い，実験にて従来のモデルを越える精度を達成した．シーングラフ生成は，画像を認識して詳細で簡潔な表現形式で記述するタスクである．具体的には，与えられた画像に映る物体と，それらの間の関係 (predicate) を表したグラフを構築する．この表現形式は，シーングラフと呼ばれている．形式的には，二つの物体と，それらの間の predicate からなる三つ組を要素とする集合として定義することができる．

我々は，このタスクにおける二つの課題に着目した．一つは，物体と predicate の相互依存の関係を考慮したシーングラフ生成の手法が必要となることである．物体とそれらの間に成り立つ predicate は互いに独立ではなく，物体のクラスが定まれば predicate のクラスの推定が容易になる一方で，predicate のクラスが定まれば物体のクラスの推定が容易になる．この相互依存性を解決するためには，組み合わせ問題を効率的に解く必要がある．もう一つは，学習データに存在しない物体クラスの組み合わせが認識時に現れ得ることである．このような場合にも，それらの間の predicate を適切に認識する必要がある．従来のシーングラフ生成モデルは，この課題に対する対処が不十分であり，predicate 認識のために極めて大きな学習データを必要とする．我々はこの二つの課題の同時解決に取り組む．

我々は，物体クラス同士の類似を捉えることで，未知の物体クラスの組の間に成り立つ predicate を推定することを目指す．未知の物体クラスの組の間に成り立つ関係は，類似した既知の物体クラスの組の間に成り立つ predicate と類似すると考えられる．そのため，物体クラス同士の類似を捉えることができれば，未知の物体クラスの組に対応できるようになると期待できる．そこで，物体クラスを離散的な記号として扱うのでなく，言語的な特徴量によって表すことを提案する．言語特徴量とは単語一つ一つに与えられる実ベクトルであり，単語同士の言語的な類似を捉えるために利用できる．これによって，物体クラス同士の類似が捉えられるようになることを期待する．

また，言語特徴量を有効に活用するためのモデルの枠組みも提案する．従来のシーングラフ生成モデルはいずれも言語特徴量を反映させる方法が明らかではない．そこで，我々は

CRF (Conditional Random Fields) の枠組みの上でニューラルネットワークのモデルを構築する．CRF では，グラフ全体を扱う大局的なモデルを，部分グラフについての局所的なモデルに分解できる．我々は局所的なモデルとして，シーングラフ中の各ノードについて物体クラスの推定を行うモデルと，各三つ組について物体クラスの組と predicate クラスの共起度を捉えるモデルを構成した．言語特徴量は，未知の物体クラスの組における predicate を推定するために導入されるため，三つ組に対応するモデルの入力として扱われる．

さらに，我々は提案するモデルにおける計算量を削減するための手法についても，提案を行う．シーングラフ中の全ての物体と predicate は互いに独立ではないため，それぞれのクラスの同時確率を扱わなければならず，計算量が大きい．これに対して我々はまず，平均場近似によって各物体と各 predicate についてのの周辺確率を近似的に求め，計算量を削減する．このときの周辺化の計算式には，各三つ組についての，物体と predicate のクラスのあらゆる組み合わせを考慮する項が含まれている．我々のモデルは計算量の大きいニューラルネットワークによって構成されているため，この項について，さらなる計算量の削減が要求される．そこで我々は，計算量を削減する手法を 2 通り提案する．一つは，三つ組のクラスの組み合わせをサンプリングすることによって，計算を近似する手法である．このとき，reparameterization trick を採用することによって微分可能なサンプリングを実現し，ニューラルネットワークの出力における誤差の伝搬を可能とする．もう一つは，モデルの一部に多重線型な層を導入し，計算の順序を工夫することで計算量を削減する手法である．多重線型な層によって，隣接した 2 ノードについてのそれぞれの特徴量と，モデルの出力の間に線形性が成り立つため，三つ組のクラスのあらゆる組み合わせについての計算を一括で行うことが可能となる．

シーングラフがアノテーションされた画像データセットによる実験の結果，我々の提案する手法が従来の手法を上回る精度を達成した．また，言語特徴量の有無について比較実験を行い，シーングラフ生成における言語特徴量の有効性を明らかにした．提案する計算削減手法についても比較実験を行なったが，これらの間に精度上の大きな差は見られなかった．多重線型な層の導入による計算量の削減はモデルの構成に制約を与えてしまうため，提案したモデルを今後発展させていくには，サンプリングによる計算量削減を扱うことが適切だと考えられる．

第1章　序論

1.1　研究背景と目的

本研究では，与えられた画像に対するシーングラフを高精度に生成する手法を提案する．

シーングラフ生成とは，画像を認識して詳細で簡潔な表現形式 (シーングラフ) で記述するタスクである．そのため，画像検索 [5] や画像質問応答 [22]，画像生成 [4] のような様々なアプリケーションへの応用が期待されている．

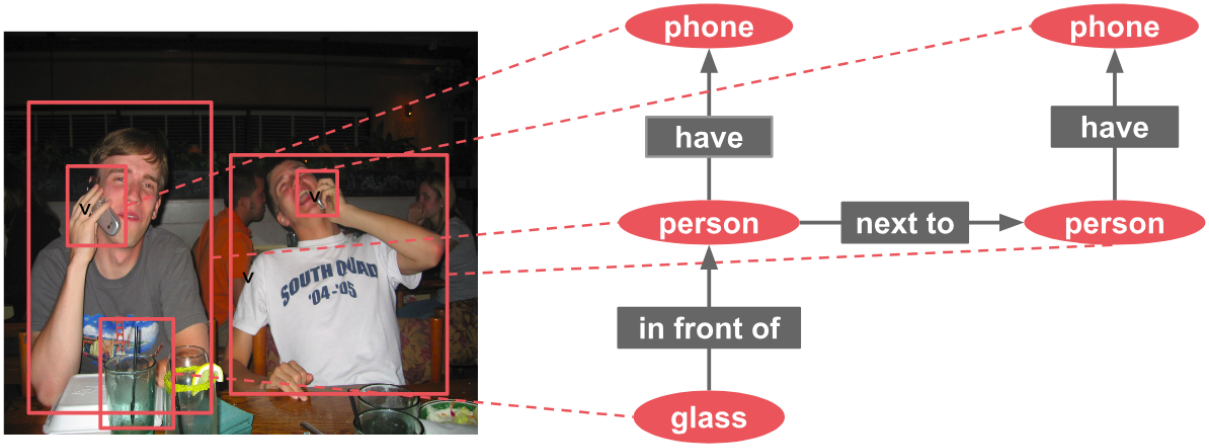

図 1.1　シーングラフの例

シーングラフは，図 1.1 に例を示すように，画像に映る物体をノードとし，それらの間に成立する関係を有向エッジとする有向グラフである．画像の意味内容をコンパクトに表現できるため，画像検索 [5] や画像質問応答 [22]，画像生成 [4] のような様々なタスクへの応用が期待されている．形式的には，$\langle$`subject`, `predicate`, `object`$\rangle$ という形式を持つ三つ組を要素とする集合として定義することができる．ここで，`subject` は主体となる物体，`object` は客体となる物体を表し，`predicate` はこれらの物体間の関係を表す．図 1.1 のシーングラフ例には，$\langle$*person*, *have*, *phone*$\rangle$ や $\langle$*glass*, *in front of*, *person*$\rangle$ などの三つ組が含まれる．

すなわち，シーングラフ生成では画像中の物体の種別を認識し，かつ，それらの間の関係 `predicate` を認識することが必要となる．

1.2 シーングラフにおける課題

シーングラフ生成における大きな課題として，次の 2 点が挙げられる．

(1) 物体のクラス認識と物体間の predicate 認識の双方を行うことが必要であること．predicate 認識においてはそれに関わる物体のクラスが正しく認識されていることが望まれるが，物体のクラス認識を決定論的に行うことは難しく，物体クラスを定めるには逆に predicate の情報が有用である．この相互依存性を解決するためには，組み合わせ問題を効率的に解く必要がある．

(2) 学習データに存在しない物体が認識時に現れる可能性があること．このような場合にも，それらの間の predicate を適切に認識する必要がある．従来のシーングラフ生成モデル [25, 23, 13, 6, 11] は，この課題に対する対処が不十分であり，predicate 認識のために極めて大きな学習データを必要とする．

本研究では，従来研究では十分に検討・達成されていない，これら二つの課題の同時解決に取り組む．

1.3 本研究のアプローチ

我々はシーングラフ生成において，画像中に未知の物体の組が現れた時に適切なシーングラフを推定することを目的とする．そのためのアプローチとして，物体のクラスを離散的な記号ではなく，それらの類似度が定義可能な連続的な言語特徴量により表現することを提案する．また，言語特徴量を有効に活用するために，CRF とニューラルネットワークを組み合わせたモデルを提案する．更に，このモデルは非常に大きい計算量を必要とするため，この計算量を削減するための計算量削減手法も併せて提案する．

我々は物体を表す単語の言語特徴量を活用することで，未知の物体の組であっても predicate を推定できると考えた．例えば，$\langle woman,\ ride,\ horse \rangle$ と $\langle man,\ ride,\ zebra \rangle$ のように，互いの subject と object が言語的に類似した物体の組は，類似した predicate を持つと考えられる．したがって，未知の物体の組の間に成り立つ predicate は，それらに類似した既知の物体の組の間に成り立つ predicate と等しいはずである．そのため，物体同士の言語的な近さを捉えることができれば，未知の物体の組の間に成り立つ predicate を推定す

ることができるはずである．我々はそのための言語特徴量として，subject と object それぞれの単語を表す分散表現 [12] を用いる．一般に，単語の分散表現とは単語に対して与えられる実ベクトルのことであり，horse と zebra のような言語的に類似した単語の組は，単語分散表現上でも近いことが知られている．この性質を利用することで，類似した物体の組に対して類似した predicate を推定することが出来る．実際に，シーングラフ生成のサブタスクである三つ組推定 [1, 11, 28, 29, 17, 9, 16, 26] では，言語特徴量の有効性が知られている [11].

言語特徴量を有効に活用するために，CRF (Conditional Random Fields) とニューラルネットワークを組み合わせたモデルを提案する．我々は，従来の研究と同様にシーングラフ生成を 3 つの段階に分け，最も重要と考えられる部分を解くためのモデルを提案する．初めに，物体である可能性が高い領域 (物体領域) を複数検出する．次に，任意のノード間が 2 つのエッジを持つように，物体領域をノードとしたグラフを構築する．このグラフの構造はシーングラフと同様であり，ノードは物体，エッジは predicate に相当する．最後に，グラフ中の全てのノードに対して物体クラスの推定を行い，全てのエッジに対して predicate クラスの推定を行うことで，シーングラフが完成される．これらの段階のうち，物体検出を除いたタスクをシーングラフ認識と呼ぶ．我々は，シーングラフ認識を解くためのモデルを，CRF の枠組みの上で構成する．CRF とは確率的グラフィカルモデルの一種であり，エネルギー関数と呼ばれる関数によって確率分布を定義する．このエネルギー関数は，グラフ中の全ての変数に対して何らかのクラスを当てはめたときの，グラフ全体の妥当性を評価するものである．またこのエネルギー関数は複数のエネルギー項から成り立っており，それぞれのエネルギー項はグラフ中のノードや部分グラフを局所的に評価する．すなわち，ノードに対応する変数や，部分グラフ中の全ての変数について何らかのクラスを当てはめた時の，ノードや部分グラフの妥当性を測るものである．我々は，グラフ中のノードと三つ組に対応する二種類のエネルギー項を設けた．ノードに対応するエネルギー項は，各物体領域に対して物体クラスの推定を行う関数である．また，三つ組に対応するエネルギー項は，subject, predicate, object ののクラスの組み合わせの妥当性を評価する関数であり，物体の共起性を捉えるとともに，predicate クラスを推定する役割を担う．この二つのエネルギー項には，ニューラルネットワークのような表現力の高い関数が必要であると考えられるため，ノードに対応するエネルギー項を CNN (Convolutional Neural Network)

で構成し，三つ組に対応するエネルギー項を MLP (Multilayer Perceptrons) によって構成する．我々は未知の物体クラスの組の間に成り立つ predicate を推定するために，三つ組に対応するエネルギー項を担う MLP の入力として，subject と object の言語特徴量を扱う．これによって，$\langle woman, horse \rangle$ と $\langle man, zebra \rangle$ のような，subject や object の言語特徴量が互いに類似した物体の組に対して，類似した predicate の推定が期待できる．

我々の提案したモデルは計算量が非常に大きいため，計算量を削減するための手法についても提案を行う．CRF によるシーングラフの認識では，グラフ中の全ての物体と predicate の同時確率分布を計算しなければならないために計算量が大きい．これに対して我々はまず，平均場近似によって各物体と各 predicate についてのの周辺確率を近似的に求め，計算量を削減する．平均場近似とは，近似的に各変数の周辺確率分布を求める手法であり，CRF の計算量を大幅に減らすことができる．このときの周辺化の計算式には，各三つ組についての，物体と predicate のクラスのあらゆる組み合わせを考慮する項が含まれている．我々のモデルは計算量の大きいニューラルネットワークによって構成されているため，この項について，さらなる計算量の削減が要求される．そこで，我々は 2 通りの計算量削減手法を提案する．一つは，三つ組のクラスの組み合わせをサンプリングすることによって，計算を近似する手法である．このとき，reparameterization trick を採用することによって微分可能なサンプリングを実現し，ニューラルネットワークの出力における誤差の伝搬を可能とする．もう一つは，モデルの一部に多重線型な層を導入し，計算の順序を工夫することで計算量を削減する手法である．多重線型な層によって，隣接した 2 ノードについてのそれぞれの特徴量と，モデルの出力の間に線形性が成り立つため，三つ組のクラスのあらゆる組み合わせについての計算を一括で行うことが可能となる．

我々は Visual Genome [7] と呼ばれる，シーングラフがアノテーションされた画像データセットを用いて実験を行う．実験では，従来のモデルと我々のモデルをシーングラフ認識の精度で比較し，我々の提案したモデルの妥当性を確かめる．また，言語特徴量の活用の有無についても比較実験を行い，言語特徴量がシーングラフ認識の精度において有効であるかどうかを明らかにする．さらに，提案した 2 通りの計算量削減手法についても，認識精度の比較を行う．

第2章　関連研究

2.1　シーングラフ生成のためのデータセット

シーングラフ生成に関する従来の研究では，主に二つの画像データセットが，評価実験のために用いられている．これらはそれぞれ，VRD (Visual Relationship Detection) データセット [11]，Visual Genome [7] と呼ばれている．これらのデータセットに含まれる全ての画像には，それぞれシーングラフがアノテーションされている．すなわち，画像中の全ての物体についての物体ラベルと，その物体を囲う矩形の物体領域がアノテーションされており，いくつかの物体の間には predicate がアノテーションされている．VRD データセットと Visual Genome の大きな違いは，保有する画像の枚数と，物体と predicate のラベルの種類の多さである．Visual Genome は VRD データセットに比べて，画像の枚数が非常に多い．また，VRD データセットは物体ラベルと predicate ラベルがある程度統一されているのに対し，Visual Genome は多様なラベルを有している．

VRD データセット [11] は 5,000 枚の画像を保持しており，100 の物体ラベルと 70 の predicate ラベルがアノテーションされている．これらの画像にアノテーションされている三つ組を全て数え上げると，その数は 37,993 であり，⟨subject，predicate，object⟩ のラベルの組み合わせは 6,672 種である．VRD データセットは，これ以前に存在した画像データセット [5, 20] に比べて，シーングラフ生成の評価実験を行うのに適切である． [5] のデータセットでは，一つの物体に対して，共起する predicate の種類が少ない．評価実験においては，物体クラスを識別することによってそれに関する predicate がほとんど特定されてしまうため，predicate の認識精度を適切に評価できない．また， [20] のデータセットは，三つ組のラベルの組み合わせの種類が非常に少ない．これらに対して VRD データセットは，一つの物体ラベルに対して様々な predicate が共起し，多様な三つ組を含むため，シーングラフ生成の評価実験を行うのに適切である．

Visual Genome [7] は 108,077 枚の画像を保持しており，それぞれの画像にはクラウドワーカーによってシーングラフがアノテーションされている．アノテーションされる物体と predicate のラベルはクラウドワーカーが自由に定めることが許されている．そのため，

Visual Genome にアノテーションされているラベルの種類は非常に多い．ただし，異なるラベルでも同一の (あるいは非常に類似した) 意味を指すものが多く含まれている．例えば，*ride* と *rides* のように変化形によって異なるものや，*person* と *man* のように包含関係にあるもの，*have* と *hold* のように同一の動作を示すものなどが存在する．シーングラフ生成の評価実験においてはこの性質が問題となることがある．シーングラフ生成の精度を評価する際には，シーングラフ生成によって推定されたシーングラフと，実際に画像にアノテーションされているシーングラフを比較する．同一の意味を持つラベルが複数あることで，二つのシーングラフの比較を困難にしてしまうため，適切な評価を行うことができなくなる．

我々は評価実験において，より多くの画像を保有する Visual Genome を採用する．ただし，データセット中に存在する物体ラベルと `predicate` ラベルのうち，100 の物体ラベルと，50 の `predicate` ラベルのみを選んで扱う．

2.2 従来のシーングラフ生成モデル

これまでのシーングラフ生成の研究には，物体認識と `predicate` 認識の相互依存関係を扱うもの [23, 13, 6] と，これらを独立に扱うもの [25] がある．

Xu ら [23] の提案した手法では，検出された物体領域をもとに構築されたグラフにおいて，ノードとエッジにそれぞれ特徴量を与え，それらを互いに伝搬させた後に，物体と `predicate` のクラスを推定する．特徴量の伝搬を行うためのモデルとして，Graph-LSTM[8] をもとに構築した Graph-GRU を用いている．シーングラフ中の全てのノードとエッジにそれぞれ GRU ユニットを設け，ノードとエッジの間で特徴量を伝搬させることで，それぞれの GRU ユニット内の中間特徴量が更新される．この操作を何度か繰り返すと，グラフ中の全ての物体領域を考慮した物体認識，`predicate` 認識が可能となる．

Newell ら [13] は，畳み込み層のみで構成されるニューラルネットワーク [15, 14] によってシーングラフ生成を行う．このニューラルネットワークは，入力された画像をダウンサンプリングして特徴量を抽出するモジュールと，これによって得られた特徴量マップをアップサンプリングして物体と `predicate` の認識を行うモジュールから成る．この構造は他のシーングラフ認識手法とは大きく異なっており，ダウンサンプリングとアップサンプリングによって複数の物体領域間で情報を伝搬させる．

Zellers ら [25] は，物体と predicate の認識を独立に行う手法を提案している．彼らは初めに物体認識のみを行い，検出された物体領域について物体クラスを特定する．その後，特定された物体クラスを用いて物体間の predicate 認識を行う．彼らのモデルは双方向 LSTM によって構成されており，各ノードに対して LSTM ユニットを設け，これらのユニット間で特徴量の伝搬を行っている．また，特徴量を伝搬させる際にノードは一列に並べられており，彼らは評価実験において様々な並べ方を比較している．それらのうち，ノードに対応する物体領域の位置によって定めた並べ方が最も良い精度を達成した．具体的には，各物体領域の中心点の，水平方向の座標にしたがってノードを並べる．従来の手法の中では最も良いシーングラフ生成精度を達成している手法である．

従来のシーングラフ生成モデルはいずれも画像特徴量を入力とした end-to-end なニューラルネットワークであり，言語特徴量を有効的に活用するための構造をもたない.. これに対し，我々は CRF の枠組みによって言語特徴量を容易に導入することを可能とする．

2.3 言語特徴量を活用した三つ組推定

シーングラフ生成のサブタスクとして，三つ組推定が盛んに研究されている [1, 11, 28, 29, 17, 9, 16, 26, 24]．三つ組推定とは，与えられた画像に映る任意の物体の組について，主体と客体を定めたときの，⟨subject，predicate，object⟩ のそれぞれのクラスを推定するタスクである．ただし，同じ物体を参照する複数の三つ組において，その物体に対して同一のクラスが推定される保証はない．また，シーングラフ生成と異なり．それぞれの三つ組が独立に推定されるため，その三つ組の外にある物体や predicate の情報を一切取り入れることができない．

言語特徴量は三つ組を推定するタスクにおいて有効であることが，従来の研究 [11] より明らかとなっている．Lu ら [11] は CNN から得た画像特徴量のみを用いた場合と，言語特徴量 [12] を加えた場合で三つ組推定の精度を比較し，言語特徴量の有効性を明らかにした．これに対し我々は，シーングラフ生成でも言語特徴量を有効に活用させることが可能であるかどうかを明らかにする．また，言語特徴量をシーングラフ生成に活用する手法を提案する．

2.4 CRF とニューラルネットワークを統合したモデル

Zheng ら [27] は，セマンティックセグメンテーションのタスクを対象として，CRF とニューラルネットワークを統合したモデルの提案を行った．また，我々と同様に，計算量を削減するために平均場近似を用いている．

セマンティックセグメンテーションとは，与えらえた画像に映る物体を全て検出し，物体の輪郭をピクセル単位で認識するタスクである．そのため，隣り合うピクセルが同一の物体に含まれているかどうかを判定する必要がある．

画像中の各ピクセルをそれぞれノードとみなすと，画像全体はノードが格子状に連結したグラフとなる．彼らは，エネルギー関数をノードに対応するエネルギー項と，隣接した 2 ノードに対応するエネルギー項で構成している．ノードに対応するエネルギー項は多層の CNN によって構成されており，隣接した 2 ピクセルに対応するエネルギー項はガウスカーネルとなっている．第 3.2 節にて述べるが，平均場近似を適用した CRF において問題となるのは，複数のノードの組に対応するエネルギー項についての計算量である．我々の手法は彼らと異なり，複数のノードの組に対応するエネルギー項を，計算量の大きいニューラルネットワークによって構成している．したがって，更なる計算量の削減が必要となるため，我々はこれを解決するための計算量削減手法を提案する．

第3章　提案手法

我々の提案手法について詳しく述べる．初めに，シーングラフ生成を行う過程を複数の段階に分けて説明し，本研究が取り組む部分を明確にする．その次に，提案するモデルの構成について述べる．最後に，提案するモデルにおける計算量を削減するための手法を詳説する．

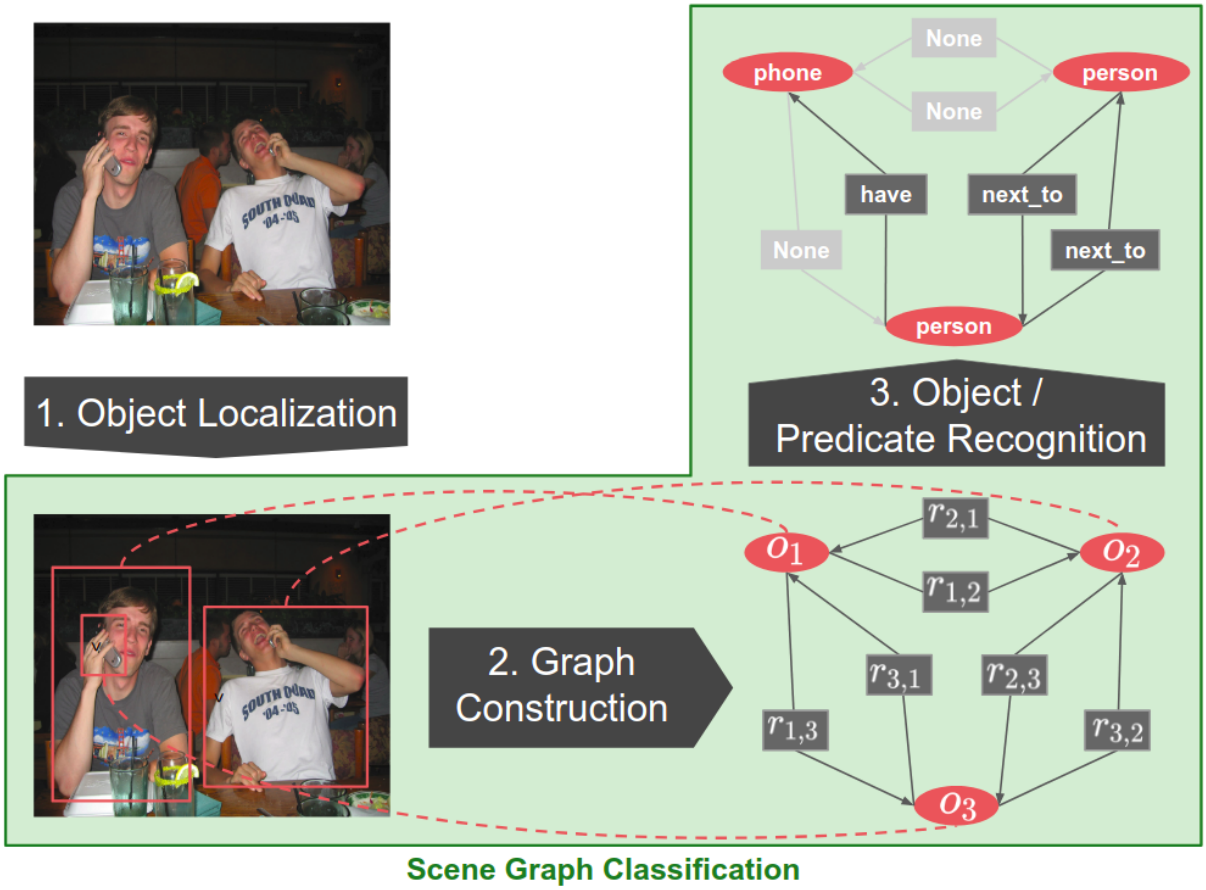

図 3.1　シーングラフ生成の流れ

図 3.1 にシーングラフ生成の流れを示す．

我々は従来の研究と同様にシーングラフ生成を三つの段階に分ける．

(1) 与えられた画像に対して，矩形の物体領域を検出する．これらの物体領域はバウンディングボックスと呼ばれる矩形であり，物体全体を囲む最小の大きさとなるように定められる．このとき，検出された物体領域の数を N と表す．

(2) 検出された物体領域候補の数だけノードをもつグラフを構築し，全ての 2 ノード間に，互いに向きが異なる有向エッジを二本設ける．このグラフはシーングラフの構造を成しており，ノードは各物体領域候補に対応する物体を表し，エッジは各物体間に成り立つ predicate を表す．また，ノード間に設けられた 2 本のエッジは，連結した 2 つのノードのどちらを subject とするかによって向きが異なる．この時点では構成したグラフ中のノードとエッジに当てはまる値が分からないため，それぞれのノードに対して，物体クラスを値にとる確率変数 $O = o_1, o_2, ..., o_N$ を設け，それぞれのエッジに対して predicate のクラスを値にとる確率変数 $R = r_{1,2}, r_{1,3}, ..., r_{N-1,N}$ を設ける．

(3) 構成したグラフ中の全ての変数 O，R に対して物体クラスと predicate クラスを与えるため，物体と predicate の識別を行う．このとき，どの物体クラスも当てはまらないようなノードや，どの predicate クラスも当てはまらないようなエッジについては，それぞれ *background* というクラスとして扱い，グラフから削除する．また，background クラスであると推定されたノードに連結するエッジも同様に削除する．

これらの段階のうち，(1) の物体領域の検出は既存の物体領域検出器 [2, 19, 10, 18] によって達成することができ，(2) のグラフの構築の方法は一意であるため，我々はシーングラフ生成から (1) と (2) とを除いた (3) の部分を研究対象とし，このタスクをシーングラフ認識と呼ぶ．すなわち，シーングラフ認識のタスクにおいては，事前に正しい物体領域が検出されていることを前提とする．

3.1　シーングラフ認識のためのモデル

我々は，シーングラフ認識を CRF の枠組みによって行う．同時確率 $p(O, R)$ は，規格化定数 Z とエネルギー関数 $E(O, R)$ を用いて以下のように定式化される．

$$p(O,R) = \frac{\exp(-E(O,R))}{Z} \tag{3.1}$$

$$Z = \sum_{o_1}\sum_{o_2}...\sum_{o_N}\sum_{r_{1,2}}\sum_{r_{1,3}}...\sum_{r_{N-1,N}} \exp(-E(O,R)) \tag{3.2}$$

$$E(O,R) = -\sum_{i}\psi_u(o_i) - \sum_{i<j}(\psi_p(o_i, r_{i,j}, o_j) + \psi_p(o_j, r_{j,i}, o_i)) \tag{3.3}$$

式中の ψ_u と ψ_p はどちらもエネルギー項である．ψ_u はグラフ中の各ノードに対応しており，それぞれの物体領域に対して物体クラスを推定する．また，ψ_p はグラフ中の各三つ組に対

応し，subject, predicate, object ののクラスの組み合わせの妥当性を評価する．例えば $\psi_u(o_i)$ は，物体領域 i の物体クラスが o_i であるかどうかを評価しており，妥当なクラスである場合は高い値を出力することが期待される．同様に $\psi_p(o_i, r_{i,j}, o_j)$ は，物体領域 i を主体，物体領域 j を客体としたときの三つ組において，subject, predicate, object のそれぞれのクラスが o_i，$r_{i,j}$，o_j であるかどうかを評価している．我々はこれらのエネルギー項を，図 3.2 と図 3.3 に示したニューラルネットワークによって構成する．

3.1.1　物体認識

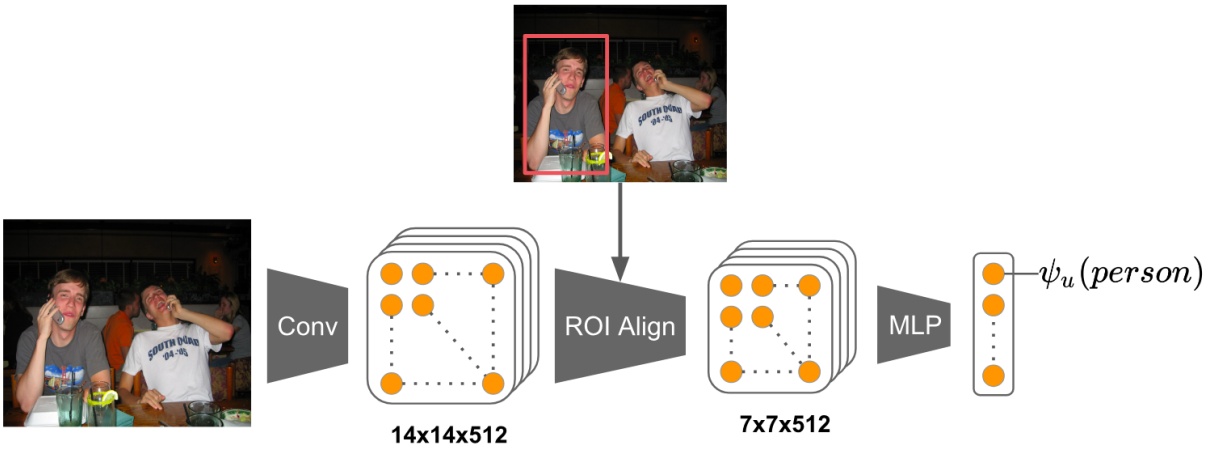

図 3.2　ノードに対応するエネルギー項 ψ_u

図 3.2 に示すように，エネルギー項 ψ_u は Conv, RoI Align, MLP の三つのモジュールによって構成されるニューラルネットワークである．Conv モジュールは，与えられた画像から画像特徴量を得るために用いられる．これに連結する RoI Align モジュールは，Conv モジュールによって得られた画像全体の特徴量から各物体領域毎の特徴量を抽出する．最後部の MLP モジュールは，各物体領域に対し，それぞれの画像特徴量を用いて物体クラスを推定する．我々は，Visual Genome [7] で事前学習された Faster-RCNN [19] から，画像特徴量抽出の役割を担っている部分を抜き出して Conv モジュールとして扱い，物体クラスの識別の役割を担っている部分を抜き出して MLP モジュールとして扱う．また，RoI Align モジュールは Mask R-CNN [2] で扱われている機構である．

Conv モジュールは 13 層の畳み込み層で構成されており，VGG 16-layer [21] から全結合層を除いたものと同様である．我々はシーングラフ認識の学習において，Conv モジュール

のパラメータを更新されないように固定する．

RoI Align モジュール [2] は，画像を畳み込み層に通して得られる特徴量マップから，任意の部分領域に相当する特徴量マップを切り出す機構である．また，切り出した特徴量マップのサイズは任意であり，入力した画像のサイズに影響されない．我々が RoI Align モジュールによって切り出す物体領域毎の特徴量マップのサイズは 7x7x512 である．

MLP モジュールは，ψ_u において唯一学習される部分であり，3 層の全結合層から構成される．このモジュールの入力は RoI Align によって得られる 7x7x512 のサイズの特徴量マップである．出力層は物体クラスと同数のユニットを持ち，それぞれのユニットはいずれかの物体クラスに対応する．その出力層は物体クラスと等しい数のユニットを持ち，それぞれのユニットはいずれかの物体クラスに対応する．シーングラフ認識ではなく，シーングラフ生成のタスクを対象とする場合は，全ての物体領域にいずれかの物体クラスが当てはまるとは限らないため，background クラスに相当するユニットを 1 つ加える．

3.1.2 predicate 認識

図 3.3 に示すように，エネルギー項 ψ_p は MLP (multilayer perceptrons) のみによって構成されるニューラルネットワークである．このニューラルネットワークは，subject と object についての 3 種類の特徴量 (言語特徴量，画像特徴量，領域特徴量) を入力とする．また，4 つの MLP に分解することができ，そのうち 3 つは入力される 3 種類の特徴量を低次元のベクトルへと圧縮する役割を担う．また，残りの 1 つは圧縮されたベクトル全てを連結したものを入力として受け，このニューラルネットワークの最終的な出力を行う．

言語特徴量は，subject，object それぞれの物体クラスを表す単語に対して得られる単語分散表現である．この単語分散表現は，Skip-gram [12] によってテキストコーパスから得られたものであり，その次元数は 300 である．ここで，我々は Wikipedia をテキストコーパスとして用いる．画像特徴量は，ψ_u の Conv モジュールと RoI Align モジュールによって得られる，subject と object それぞれの物体領域についての特徴量マップである．また，図 3.4 に示すように，領域特徴量は物体領域間の相対位置と被覆面積を 4 次元で表した実ベクトルである これらの特徴量を圧縮する MLP はどれも 2 層の全結合層から構成される．また，画像特徴量を圧縮する MLP の構造は ψ_u の MLP モジュールの前段の 2 層と等しく，パラメータの初期値についても同様である．

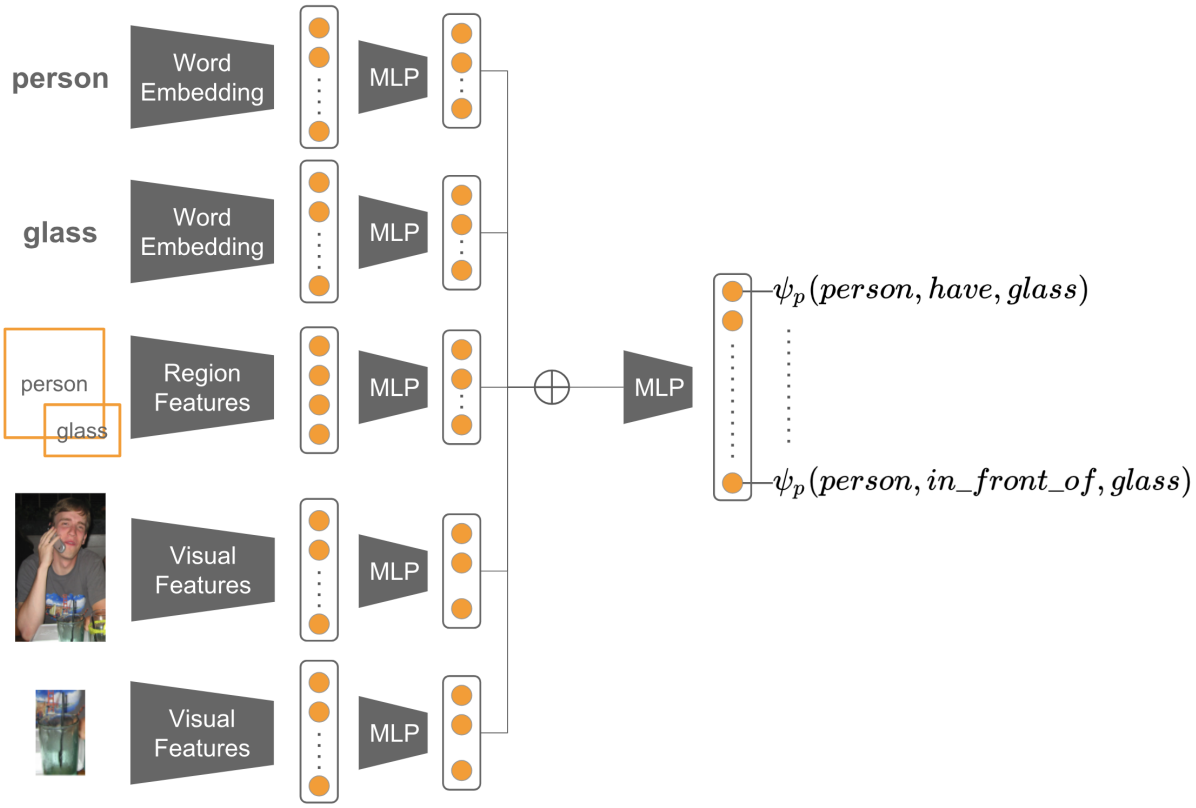

図 3.3　三つ組に対応するエネルギー項 ψ_p

ψ_p の出力層は predicate と等しい数のユニットを持ち，それぞれのユニットはいずれかの predicate クラスに対応する．ただし，全ての物体間にいずれかの predicate クラスが当てはまるとは限らないため，background クラスに相当するユニットを 1 つ加える．

3.2　シーングラフ認識の計算手法

我々は二つのエネルギー項 ψ_u，ψ_p によって同時確率 $p(O, R)$ を計算するが，式 3.1 における規格化定数 Z の計算量は明らかに大きく，計算が困難である．

これに対して我々はまず，平均場近似を適用して計算量を削減する．平均場近似では，各変数についてそれぞれ周辺確率分布を求め，それらの積によって同時確率分布を近似する．シーングラフ認識では，グラフ中のノードの一つ一つに対して物体クラスの周辺確率分布を求め，エッジの一つ一つに対して predicate クラスの周辺確率分布を求めることに相当

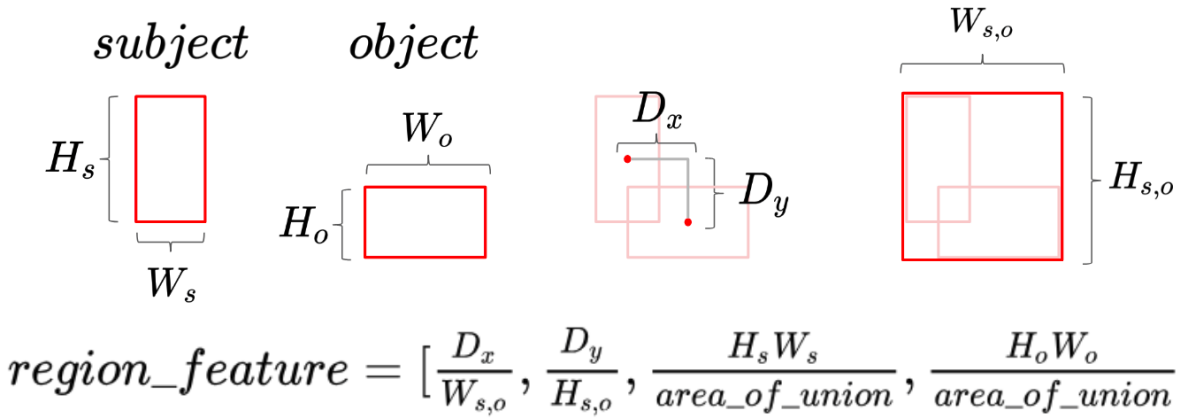

図 3.4　領域特徴量の獲得方法

する．具体的には，近似した同時確率 $q(O,R)$ は以下の式で表される．

$$q(O,R) = \prod_i q(o_i) \cdot \prod_{i<j} q(r_{i,j}) q(r_{j,i}) \tag{3.4}$$

これにより規格化定数 Z の計算を避けることができるため，CRF の計算量を大幅に減らすことができる．

式 3.4 と式 3.1 の確率分布間のカルバック・ライブラー・ダイバージェンスを最小化することによって，これらの近似を行う．以下に，最小化すべきカルバック・ライブラー・ダイバージェンスを示す．

$$KL(q\|p) = \sum_O \sum_R \prod_i q(o_i) \cdot \prod_{i<j} q(r_{i,j}) q(r_{j,i}) \frac{\prod_i q(o_i) \cdot \prod_{i<j} q(r_{i,j}) q(r_{j,i})}{\exp(-E(O,R))} \tag{3.5}$$

これを，各周辺確率に関する拘束条件 $\sum_o q(o) = 1$，$\sum_r q(r) = 1$ のもとで，$q(o)$ と $q(r)$ について微分することで以下の更新式を得る．

$$q^0(o_i) \propto \exp(\psi_u(o_i)) \tag{3.6}$$

$$q^{t+1}(o_i) \propto \exp(\psi_u(o_i) + \sum_{j \neq i} e_{obj}(i,j)) \tag{3.7}$$

$$\begin{aligned} e_{obj}(i,j) = &\sum_{o_j, r_{i,j}} q^t(o_j) q^t(r_{i,j}) \psi_p(o_i, r_{i,j}, o_j) \\ &+ \sum_{o_j . r_{j,i}} q^t(o_j) q^t(r_{j,i}) \psi_p(o_j, r_{j,i}, o_i) \end{aligned} \tag{3.8}$$

$$q^{t+1}(r_{i,j}) \propto \exp(e_{pred}(i,j)) \tag{3.9}$$

$$e_{pred}(i,j) = \sum_{o_i,o_j} q^t(o_i) q^t(o_j) \psi_p(o_i, r_{i,j}, o_j) \tag{3.10}$$

これらの計算を T 回行った後に得られた周辺確率 $q^T(o_i)$，$q^T(r_{i,j})$ を用いて，以下のように損失関数を定義できる．

$$L(O,R) = -\sum_i \log\ q^T(o_i) - \lambda \sum_{i<j} (\log\ q^T(r_{i,j}) + \log\ q^T(r_{j,i})) \tag{3.11}$$

この損失関数は物体クラスと `predicate` クラスについての交差エントロピー関数であり，ハイパーパラメータ λ は物体認識と `predicate` 認識の重みを調整する値となる．我々は，$T = 2$，$\lambda = 1.0$ とした．

以上のように平均場近似は計算量を大幅に減らすが，式 3.8 と式 3.10 において ψ_p の計算回数が非常に多く，依然として計算量が大きいことが分かる．我々は，この計算量を更に削減するために 2 通りの計算量削減手法を提案する．一つは，物体クラス，`predicate` クラスについての期待値がとられていることに注目し，それぞれのクラスをサンプリングすることによって ψ_p の計算回数を減らす手法である．もう一つは，ψ_p のニューラルネットワークの一部を多重線型な層で置き換えることで計算量の削減を行う手法である．

3.2.1　サンプリングによる計算量の削減

式 3.8 と式 3.10 において ψ_p の期待値がとられていることに着目し，サンプリングによる計算回数の削減を行う．

$$\begin{aligned} e_{obj}(i,j) =& E_{q^t(o_j),q^t(r_{i,j})}[\psi_p(o_i, r_{i,j}, o_j)] \\ &+ E_{q^t(o_j),q^t(r_{j,i})}[\psi_p(o_j, r_{j,i}, o_i)] \\ \simeq& \frac{1}{S_i S_{i,j}} \sum_{\substack{o_j \sim q^t(o_j) \\ r_{i,j} \sim q^t(r_{i,j})}} \psi_p(o_i, r_{i,j}, o_j) \\ &+ \frac{1}{S_i S_{j,i}} \sum_{\substack{o_j \sim q^t(o_j) \\ r_{j,i} \sim q^t(r_{j,i})}} \psi_p(o_j, r_{j,i}, o_i) \end{aligned} \tag{3.12}$$

$$\begin{aligned} e_{pred}(i,j) =& \quad E_{q^t(o_i),q^t(o_j)}[\psi_p(o_i, r_{i,j}, o_j)] \\ \simeq& \quad \frac{1}{S_i S_j} \sum_{\substack{o_i \sim q^t(o_i) \\ o_j \sim q^t(o_j)}} \psi_p(o_i, r_{i,j}, o_j) \end{aligned} \tag{3.13}$$

$$\tag{3.14}$$

式中の S_i と S_j は物体クラスのサンプリング数を表し，$S_{i,j}$ と $S_{j,i}$ は predicate クラスのサンプリング数を表す．更に我々は，サンプリングの処理を微分可能な計算とするために Gumbel-Softmax による reparameterization trick [3] を採用した．したがって，我々は以下の計算によって各変数についての周辺分布を推定する．

$$\begin{aligned} e_{obj}(i,j) =& \sum_{o_j, r_{i,j}} z^t(o_j) z^t(r_{i,j}) \psi_p(o_i, r_{i,j}, o_j) \\ &+ \sum_{o_j . r_{j,i}} z^t(o_j) z^t(r_{j,i}) \psi_p(o_j, r_{j,i}, o_i) \\ \simeq& \frac{1}{S_i S_{i,j}} \sum_{\substack{o_j \sim z^t(o_j) \\ r_{i,j} \sim z^t(r_{i,j})}} \psi_p(o_i, r_{i,j}, o_j) \\ &+ \frac{1}{S_i S_{j,i}} \sum_{\substack{o_j \sim z^t(o_j) \\ r_{j,i} \sim z^t(r_{j,i})}} \psi_p(o_j, r_{j,i}, o_i) \end{aligned} \tag{3.15}$$

$$\begin{aligned} e_{pred}(i,j) =& \sum_{o_i, o_j} z^t(o_i) z^t(o_j) \psi_p(o_i, r_{i,j}, o_j) \\ \simeq& \frac{1}{S_i S_j} \sum_{\substack{o_i \sim z^t(o_i) \\ o_j \sim z^t(o_j)}} \psi_p(o_i, r_{i,j}, o_j) \end{aligned} \tag{3.16}$$

$$z^t(x) = \frac{\exp((\log(q^t(x)) + g(x))/\tau)}{\sum_y \exp((\log(q^t(y)) + g(y))/\tau)} \tag{3.17}$$

$$g(x) = -\log(-\log(u)) \tag{3.18}$$

$$u = \mathrm{Uniform}(0, 1) \tag{3.19}$$

我々はサンプリング数を $S_i = S_j = 3$，$S_{i,j} = S_{j,i} = 30$ と設定し，温度パラメータを $\tau = 0.1$ とした．

3.2.2　多重線型な層による計算量の削減

ψ_p のニューラルネットワークの構成に多重線型な層 (Multilinear layer) を導入することによって，計算量を削減する．図 3.5 に，多重線型な層を含む ψ_p のニューラルネットワークの構成を示す．多重線型な層とは，任意の数の実ベクトルを入力することができる，多重線型な関数である．例えば，図 3.6 に示すように，2 つのベクトル $x \in \mathcal{R}^{\mathcal{A}}, y \in \mathcal{R}^{\mathcal{B}}$ を入力した場合，多重線型な層の出力 $m \in \mathcal{R}^{\mathcal{C}}$ の c 次元目の値 w_c は重み $w \in \mathcal{R}^{\mathcal{A} \times \mathcal{B} \times \mathcal{C}}$ を用いて

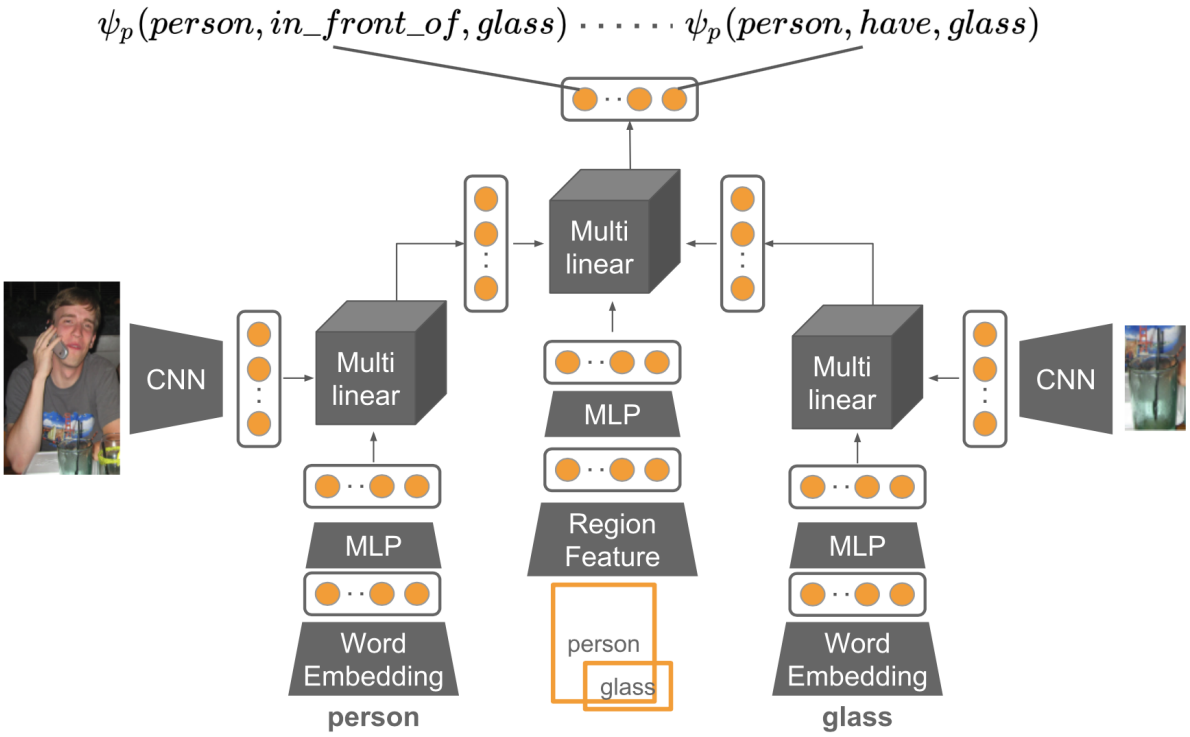

図 3.5　三つ組に対応するエネルギー項 ψ_p

以下のように表される．

$$\begin{aligned} m_c =& Multilinear(x, y) \\ =& \sum_a^A \sum_b^B w_{abc} x_a y_b \end{aligned} \tag{3.20}$$

同様に，3 つのベクトル $x \in \mathcal{R}^{\mathcal{A}}, y \in \mathcal{R}^{\mathcal{B}}, z \in \mathcal{R}^{\mathcal{C}}$ を入力とする場合，多重線型な層の出力 $m \in \mathcal{R}^{\mathcal{D}}$ の d 次元目の値 w_d は重み $w \in \mathcal{R}^{\mathcal{A}\times\mathcal{B}\times\mathcal{C}\times\mathcal{D}}$ を用いて以下のように表される．

$$\begin{aligned} m_d =& Multilinear(x, y, z) \\ =& \sum_a^A \sum_b^B \sum_c^C w_{abcd} x_a y_b z_c \end{aligned} \tag{3.21}$$

この多重線型性を有効に活用することで，式 3.8 と 3.10 の計算量を大きく削減することができる．式 3.8 と式 3.10 では，周辺化を行うために，隣接するノードについて全ての物体クラスを考慮している．その際，物体クラス毎に言語特徴量が異なるため，入力する特徴量を変えて何度も ψ_p の計算を行うことになる．ここで，ψ_p の一部に多重線型な層を導入すると，多重線型性の性質から，言語特徴量を全ての物体クラスについて足し合わせたベクトルを一度 ψ_p に入力するだけで，周辺化の計算を行うことができるようになる．例えば，式 3.8 は以下のように計算できる．. 式中の $f(x)$ は物体クラス x についての言語特徴量，$g(i, j)$ は物体ペア (i, j) に対する領域特徴量，h_{sbj} と h_{obj} はそれぞれ `subject` と `object`

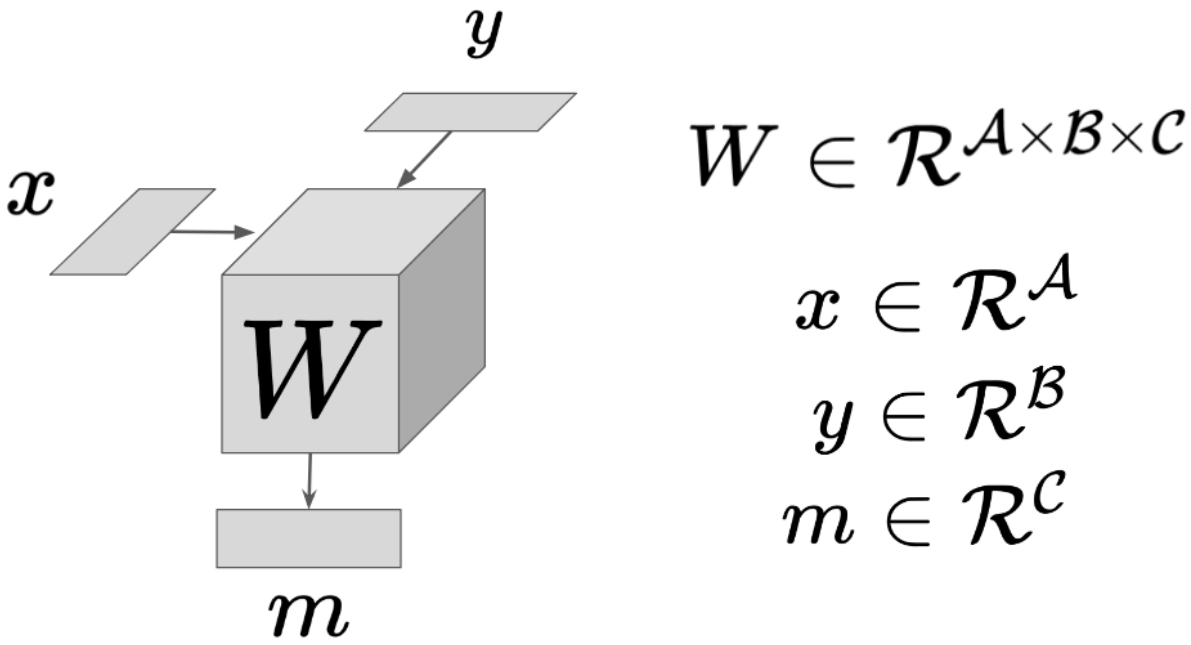

図 3.6　多重線型な層を導入した ψ_p

に対する画像特徴量である．

$$
\begin{aligned}
&\sum_{o_j, r_{i,j}} q^t(o_j) q^t(r_{i,j}) \psi_p(o_i, r_{i,j}, o_j) \\
=&\sum_{o_j, r_{i,j}} q^t(o_j) q^t(r_{i,j}) Multilinear(Multilinear(F(o_i), H_{sbj}), G(i,j), Multilinear(F(o_j), H_{obj})) \\
=&\sum_{r_{i,j}} q^t(r_{i,j}) Multilinear(Multilinear(F(o_i), H_{sbj}), G(i,j), Multilinear(\sum_{o_j} q^t(o_j) F(o_j), H_{obj}))
\end{aligned}
$$

$$
\begin{gathered}
F(o_i) = MLP(f(o_i)), \quad F(o_j) = MLP(f(o_j)) \\
G(i,j) = MLP(g(i,j)) \\
H_{sbj} = MLP(h_{sbj}), \quad H_{obj} = MLP(h_{obj})
\end{gathered}
$$

同様に，式 3.10 は以下のように計算できる．

$$
\begin{aligned}
&\sum_{o_i, o_j} q^t(o_i) q^t(o_j) \psi_p(o_i, r_{i,j}, o_j) \\
=&\sum_{o_i, o_j} q^t(o_i) q^t(o_j) Multilinear(Multilinear(F(o_i), H_{sbj}), G(i,j), Multilinear(F(o_j), H_{obj})) \\
=&Multilinear(Multilinear(\sum_{o_i} q^t(o_i) F(o_i), H_{sbj}), G(i,j), Multilinear(\sum_{o_j} q^t(o_j) F(o_j), H_{obj}))
\end{aligned}
$$

以上のように計算の順序を工夫することによって，計算量を削減することができる．

ただし，多重線型な層の導入による計算量の削減はモデルの構成に制約を与えてしまうという問題がある．また，多重線型な層のパラメータのサイズは，入力するベクトルの数が増えるに従って劇的に増大するため，特徴量の数についても制約が存在する．これらの理由から，実験において提案する 2 つの計算量削減手法を比較し，両者の精度が等しい場合は，サンプリングによる計算量削減が適切な手法であると考えられる．

第4章　シーングラフ認識精度の評価実験と考察

4.1　実験設定

我々は画像データセットとして Visual Genome [7] を用いて評価実験を行った．Visual Genome の保持する 108,077 枚の画像データセットのうち，75,631 枚の画像を学習用データセットとし，残りの 32,319 枚を評価用データセットとした．またこれらの画像にアノテーションされている物体名称と predicate のうち，150 の物体クラスと 50 の predicate のみを選んで扱う．この実験設定は [25] と全く同様である．

シーングラフ認識精度の評価のために，Recall@K [11] と呼ばれる評価手法を用いる．これは，推定されたシーングラフを三つ組の集合に分解し，それらのうち上位 K 件の三つ組に対して Recall を測るものである．また，我々の対象としているタスクはシーングラフ認識であるが，本実験では物体認識の精度についても評価を行う．これは，三つ組に対応するエネルギー項によって物体認識の精度が向上する可能性があるためである．

4.2　比較モデル

我々は，シーングラフ認識のタスクにおいて，我々の提案モデル (Ours) と 3 つのモデルを比較した．比較対象のモデルは全て，第 2.2 節で述べた既存のモデル (ASSOCIATIVE EMBEDDING，MESSAGE PASSING，MOTIFNET) である．また，物体認識のタスクにおいて，ψ_u のみを用いて物体認識を行うモデル CNN only の認識精度を確認する．我々の提案するモデルをこのモデルと比較することによって，CRF によって物体認識精度が向上するかどうかを調べることができる．

また，提案するモデルに対して 2 通りの計算量削減手法を適用し，サンプリングによる計算近似を用いたモデルを Ours（Sampling），ψ_p に多重線型な層を導入したモデルを Ours（Multilinear）と表す．シーングラフ認識，物体認識のいずれにおいても，これらの計算手法について比較を行う．

さらに，提案モデルにおける言語特徴量の有効性を評価するため，単語分散表現を one-hot 表現に置き換えたモデルとの比較を行う．one-hot 表現は物体クラスと同数の次元を持つベクトルで，各次元はいずれかの物体クラスに対応する．物体クラスを表すために，それらのうち 1 つだけ値が 1 となり，その他全ての値が 0 となる．したがって one-hoth 表現は物体クラスを離散的に表したものであり，単語分散表現の有効性を確認するために比較するべき表現である．我々は，単語分散表現を扱うモデルを Ours（Skip-gram）と表し，one-hot 表現を扱うモデルを Ours（one-hot）と表す．

4.3　実験項目

我々は，以下の 3 項目の実験を行った．

物体認識精度とシーングラフ認識精度の比較　第 4.2 節で述べたモデルにおいて，認識精度の比較を行う．また，言語特徴量の有効性を明らかにし，提案する計算量削減手法についても比較を行う．

物体と predicate の相互依存関係の重要性の確認　MOTIFNET [25] で用いられている，物体認識を特定した後に predicate を認識するパイプライン方式を我々のモデルに採用し，シーングラフ認識精度を評価する．物体と predicate を同時に認識する場合とパイプライン方式を比較することで，物体と predicate の相互依存関係がシーングラフ認識において有用かどうかを確認する．我々は，パイプライン方式を採用した提案モデルを Pipeline と表す．

既存モデルに対する言語特徴量の有効性の確認　既存のシーングラフ認識モデルに対して言語特徴量を導入した場合のシーングラフ認識精度を評価する．既存モデルはいずれも，物体クラスを識別するために，物体クラスと同数のユニットを持つ出力層からの出力に対して softmax 関数を用いる．我々は言語特徴量を導入するため，この出力層から言語特徴量と同じ次元数のベクトルを出力するようにし，出力されたベクトルと各物体クラスを表す言語特徴量との内積に対して softmax 関数を用いる．これによって類似した物体クラスを区別せず扱えるようになるため，画像中に未知の物体の組が現れた場合にも適切なシーングラフが推定できることが期待できる．

4.4　結果と考察

表 4.1　物体認識精度とシーングラフ認識精度における各モデルの比較結果

	物体認識	シーングラフ認識	
モデル	識別率	R@50	R@100
CNN only	64.7	-	-
MESSAGE PASSING [23]	-	34.6	35.4
ASSOCIATIVE EMBEDDING [13]	-	21.8	22.6
MOTIFNET [25]	-	35.8	36.5
Ours (Sampling, one-hot)	65.1	35.5	36.5
Ours (Sampling, Skip-gram)	66.6	36.1	36.8
Ours (Multilinear, one-hot)	65.2	35.3	36.0
Ours (Multilinear, Skip-gram)	**66.7**	**36.2**	**36.9**

表 4.1 に物体認識精度とシーングラフ認識精度における各モデルの比較結果の結果を示す．この表から，物体認識において全ての提案モデルが CNN only の認識精度を上回って

表 4.2　パイプライン方式を採用した提案モデルのシーングラフ認識精度

	シーングラフ認識	
モデル	R@50	R@100
Ours (Sampling, Skip-gram)	36.1	36.8
Ours (Multilinear, Skip-gram)	**36.2**	**36.9**
Pipeline (Sampling, Skip-gram)	35.9	36.7
Pipeline (Multilinear, Skip-gram)	36.0	36.7

表 4.3　言語特徴量を導入した既存モデルのシーングラフ認識精度

	シーングラフ認識	
モデル	R@50	R@100
MESSAGE PASSING [23]	34.6	35.4
MESSAGE PASSING (Skip-gram) [23]	32.4	33.6
MOTIFNET [25]	35.8	36.5
MOTIFNET (Skip-gram) [25]	30.3	31.2
Ours (Sampling, Skip-gram)	36.1	36.8
Ours (Multilinear, Skip-gram)	**36.2**	**36.9**

いることが確認できる．このことから，我々の提案モデルが物体認識においても有用であることがわかる．また，言語特徴量を利用している Ours (Sampling, Skip-gram) と Ours (Multilinear, Skip-gram) がそれぞれ Ours (Sampling, one-hot) と Ours (Multilinear, one-hot) に対してより良い物体認識精度を得ていることが確認できる．したがって，言語特徴量の活用が物体認識精度に貢献することが明らかとなった．

また，シーングラフ認識においても言語特徴量を利用した Ours (Sampling, Skip-gram) と Ours (Multilinear, Skip-gram) がそれぞれ Ours (Sampling, one-hot) と Ours (Multilinear one-hot) に対してより良い精度を得ていることがわかる．言語特徴量が物体認識のみならずシーングラフ認識においても有用であることが確認できた．更に，Ours (Sampling, Skip-gram) と Ours (Multilinear, Skip-gram) は既存のモデルより良いシーングラフ認識精度を得ており，我々の提案するモデルはシーングラフ認識において妥当なモデルであると考えられる．我々が提案する 2 通りの計算量削減手法を比較すると，Ours (Sampling, Skip-gram) と Ours (Multilinear, Skip-gram)，Ours (Sampling, one-hot) と Ours (Multilinear, one-hot) が互いにほぼ同等の精度を示していることが確認できる．そのため，これらの計算量削減手法に精度上の大きな差は無いと考えられる．

表 4.2 にパイプライン方式を採用した提案モデルのシーングラフ認識精度を示す．この表から，物体と predicate の相互依存関係を考慮したモデル (Ours (Sampling, Skip-gram), Ours (Multilinear, Skip-gram)) が，パイプライン方式を採用した提案モデル (Pipeline (Sampling, Skip-gram), Pipeline (Multilinear, Skip-gram)) に比べて高いシーング

ラフ認識精度を達成していることがわかる．ただし，その差は大きくないため，シーングラフ認識における物体認識と predicate 認識の相互依存関係の有用性については検討する余地がある．

表 4.3 に言語特徴量を導入した既存モデルのシーングラフ認識精度を示す．従来のモデル (MESSAGE PASSING，MOTIFNET) はいずれも言語特徴量を採用することで認識精度が大きく下がっている．モデルの出力したベクトルと物体クラスのラベルを表す言語特徴量の積が物体クラスの確率分布を定めるため，言語特徴量において類似した物体クラスは，互いに類似した確率を与えられてしまう．このことがシーングラフ認識の精度の悪化に影響していると考えられる．

第5章　結論

本稿では，画像中に未知の物体の組が出現したときのシーングラフ認識を適切に行うため，言語特徴量の活用を提案した．また，言語特徴量を有効に活用するためにCRFとニューラルネットワークを統合したモデルを提案した．さらに，我々の提案したモデルにおける計算量の問題を克服するための計算量削減手法を2通り提案した．1つは，周辺化の際に考慮する物体とpredicateのクラスの組み合わせのパターンを減らして計算を近似する手法であり，もう1つは，モデルの一部に線形性を導入することによって計算量を減らす手法である．Visual Genomeをデータセットとして用いた実験では，提案したシーングラフ認識モデルに言語特徴量を活用することで，従来提案されてきたモデルを上回る精度を達成した．また，言語特徴量が物体認識，シーングラフ認識において有効であることが明らかとなった．提案した2通りの計算量削減手法についての比較実験では，これらが同等の性能を示すことが確認された．多重線型な層の導入による計算量の削減はモデルの構成に制約を与えてしまうため，提案したモデルを今後発展させていくには，サンプリングによる計算量削減を扱うことが適切だと考えられる．

我々が提案した枠組みはCRFに基づいており，そのエネルギー関数は物体推定を行うエネルギー項 (ψ_u) と三つ組の妥当性を捉えるエネルギー項 (ψ_p) によって成り立っている．それぞれのエネルギー項の構成を変更することが容易であるため，より複雑なニューラルネットワークを用いることが可能となる．三つ組推定のタスクでは様々なニューラルネットワークによるモデルが提案されているため，我々のシーングラフ認識モデルにそれらを統合することで，さらなる認識精度の向上が期待できる．今後我々はこの枠組みの上で，より良いモデルの構成を探っていく．

また，どのような物体， predicateについて単語分散表現が有効に働くのか分析を行なっていく．さらに，Skip-gram以外の手法による単語分散表現の獲得について検討する．単語分散表現以外の，テキストコーパスから物体とpredicateの共起の情報を得るなどの手段についても調査を進めていく．

謝辞

本研究を進めるにあたり，多くの方々に様々なご助力を賜りました．

林良彦教授には大変多くのご指導を頂きました．日々の議論では，常に鋭いご意見を賜りました．また，論文の執筆や資料作成などにおいても，とても多くのお時間を頂きました．心より感謝申し上げます．

小林哲則教授には研究生活を強く支えて頂きました．また，本質をついた意見を頂けたおかげで，研究の方針を定めることができました．深く感謝致します．

小林哲則研究室，小川哲司研究室に所属する学生の皆様には，様々な視点から多くのアイデアを頂きました．また，日頃から皆様と議論させて頂き，とても楽しい研究生活を送ることができました．深く感謝致します．

最後に，研究のみならず様々な面でお世話になりました小林哲則研究室，小川哲司研究室の皆様に深く感謝申し上げます．

参考文献

[1] Bo Dai, Yuqi Zhang, and Dahua Lin. Detecting visual relationships with deep relational networks. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pp. 3298–3308. IEEE, 2017.

[2] Kaiming He, et al. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 2980–2988. IEEE, 2017.

[3] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

[4] Justin Johnson, Agrim Gupta, and Li Fei-Fei. Image generation from scene graphs. *arXiv preprint arXiv:1804.01622*, 2018.

[5] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3668–3678, 2015.

[6] Matthew Klawonn and Eric Heim. Generating triples with adversarial networks for scene graph construction. *arXiv preprint arXiv:1802.02598*, 2018.

[7] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, Vol. 123, No. 1, pp. 32–73, 2017.

[8] Xiaodan Liang, Xiaohui Shen, Jiashi Feng, Liang Lin, and Shuicheng Yan. Semantic object parsing with graph lstm. In *European Conference on Computer Vision*, pp. 125–143. Springer, 2016.

[9] Wentong Liao, Lin Shuai, Bodo Rosenhahn, and Michael Ying Yang. Natural language guided visual relationship detection. *arXiv preprint arXiv:1711.06032*, 2017.

[10] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pp. 21–37. Springer, 2016.

[11] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, pp. 852–869. Springer, 2016.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[13] Alejandro Newell and Jia Deng. Pixels to graphs by associative embedding. In *Advances in neural information processing systems*, pp. 2171–2180, 2017.

[14] Alejandro Newell, Zhiao Huang, and Jia Deng. Associative embedding: End-to-end learning for joint detection and grouping. In *Advances in Neural Information Processing Systems*, pp. 2277–2287, 2017.

[15] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, pp. 483–499. Springer, 2016.

[16] Khoi Nguyen. Relational networks for visual relationship detection in images. 2017.

[17] François Plesse, Alexandru Ginsca, Bertrand Delezoide, and Françoise Prêteux. Visual relationship detection based on guided proposals and semantic knowledge distillation. *arXiv preprint arXiv:1805.10802*, 2018.

[18] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. *arXiv preprint*, 2017.

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.

[20] Mohammad Amin Sadeghi and Ali Farhadi. Recognition using visual phrases. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pp. 1745–1752. IEEE, 2011.

[21] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[22] Damien Teney, Lingqiao Liu, and Anton van den Hengel. Graph-structured representations for visual question answering. *CoRR, abs/1609.05600*, Vol. 3, , 2016.

[23] Danfei Xu, Yuke Zhu, Christopher B Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2, 2017.

[24] Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Visual relationship detection with internal and external linguistic knowledge distillation. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[25] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. *arXiv preprint arXiv:1711.06640*, 2017.

[26] Hanwang Zhang, Zawlin Kyaw, Shih-Fu Chang, and Tat-Seng Chua. Visual translation embedding network for visual relation detection. In *CVPR*, Vol. 1, p. 5, 2017.

[27] Shuai Zheng, et al. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 1529–1537, 2015.

[28] Yaohui Zhu and Shuqiang Jiang. Deep structured learning for visual relationship detection. 2018.

[29] Bohan Zhuang, Lingqiao Liu, Chunhua Shen, and Ian Reid. Towards context-aware interaction recognition for visual relationship detection. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pp. 589–598. IEEE, 2017.